PHISHING WEBSITE DETECTION

NUR SHOLIHAH BINTI ZAINI

Bachelor of Computer Science
(Computer System and Networking)

UNIVERSITI MALAYSIA PAHANG

SUPERVISOR'S DECLARATION

I hereby declare that I have checked this thesis and in my opinion, this thesis is adequate in terms of scope and quality for the award of the degree of Computer Science in Networking.

_____

(Supervisor's Signature)

Full Name : DR MOHD FAIZAL BIN AB RAZAK

Position : SENIOR LECTURER

Date : 7 JANUARY 2019

STUDENT'S DECLARATION

I hereby declare that the work in this thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at Universiti Malaysia Pahang or any other institutions.

_____

(Student's Signature)

Full Name : NUR SHOLIHAH BINTI ZAINI

ID Number : CA15080

Date : 7 JANUARY 2019

PHISHING WEBSITE DETECTION

NUR SHOLIHAH BINTI ZAINI

Thesis submitted in fulfillment of the requirements
for the award of the degree of
Bachelor of Computer Science

Faculty of Computer System and Software Engineering

UNIVERSITI MALAYSIA PAHANG

DECEMBER 2018

ACKNOWLEDGEMENTS

In the accomplishment of this research successfully, many people have best owned upon me their blessings and the heart pledged support, this time I am utilizing to thank all the people who have been concerned about this research.

Primarily I would to thank Allah for being able to complete this research with success. Then I would like to thank my supervisor, Dr Abdul Ghani Ali Ahmed and Dr Mohd Faizal bin Ab Razak whose valuable guidance has been the ones that helped me patch this research and make it success. Their suggestion and instruction have served as major contributors towards the completion of the research.

Then I would like to thank my parent, Zaini bin Taib and Noridah binti Ismail for giving me support both morally and finance. Next, I would like to express my gratitude to my friends, Ibrahim bin Othman, Nurul Farah Anisah binti Jenimen, Ong Vienna Lee and my roomates who have been helped me with their valuable suggestion and moral support in various phases of the completion of the research.

Last but not least I would like to thank the entire people who directly or indirectly contribute on completing this research.

ABSTRAK

Internet telah menjadi sebahagian daripada aktiviti sosial dan kewangan harian kami. Internet adalah penting bukan sahaja untuk pengguna individu, tetapi juga untuk organisasi, lebih lebih lagi sebagai organisasi yang menawarkan perdagangan dalam talian dapat memperoleh kelebihan daya saing dengan menawarkan pelbagai perkhidmatan kepada pelanggan global. Internet memungkinkan untuk mencapai pelanggan di seluruh dunia tanpa sekatan pasaran dan dengan e-dagang yang berkesan. Akibatnya, bilangan pelanggan yang menggunakan Internet untuk membuat pembelian mereka meningkat dengan ketara. Beratus-ratus juta dolar dipindahkan setiap hari melalui internet. Jumlah wang ini menarik perhatian penjenayah siber untuk menjalankan aktiviti haram mereka. Oleh itu, pengguna Internet mungkin terdedah kepada pelbagai jenis ancaman web yang boleh menyebabkan kerugian kewangan, penipuan kad kredit, kehilangan data peribadi, menjejaskan reputasi organisasi dan kehilangan kepercayaan dalam perkhidmatan e-dagang dan perbankan dalam talian oleh pelanggan. Oleh itu, kesesuaian internet untuk urus niaga komersial akan dipersoalkan. Phishing dianggap sebagai ancaman web yang didefinisikan sebagai seni penyamaran sebagai laman web sebenar untuk mendapatkan nama pengguna, kata laluan dan butiran kad kredit. Dalam kajian ini, fenomena Phishing akan dibincangkan secara terperinci. Di samping itu, kami membentangkan kajian mengenai cara penyelidikan berkenaan topik ini. Tambahan pula, penyelidikan ini bertujuan untuk mengenal pasti perkembangan terkini dalam phishing dan langkah berjaga-jaga, serta menjalankan kajian dan penilaian komprehensif terhadap penyelidikan ini untuk menutup jurang yang masih wujud dalam topic ini. Penyelidikan ini tertumpu terutamanya pada kaedah pengesanan phishing data berasaskan web, tidak tertumpu kepada kaedah pengesanan berasaskan e-mel.

ABSTRACT

The Internet has become an integral part of our daily social and financial activities. The Internet is important not only for individual users, but also for organizations, as organizations that offer online commerce can gain a competitive advantage by serving global customers. The Internet makes it possible to reach customers all over the world without market restrictions and with effective e-commerce. As a result, the number of customers using the Internet to make their purchases is increasing significantly. Hundreds of millions of dollars are transferred every day over the internet. This amount of money tempted the fraudsters to carry out their illegal activities. Therefore, Internet users may be vulnerable to various types of web threats that may cause financial harm, credit card fraud, loss of personal data, potential damage to brand reputation, loss of trust in e - commerce and online banking by customers. Hence, the suitability of the internet for commercial transactions is questionable. Phishing is considered a form of web threats that is defined as the art of impersonating a legit website to obtain usernames, passwords and credit card details. In this study, the phishing phenomena will be discussed in detail. In addition, we present a study on the state of research on the topic. Furthermore, we aim to identify the current developments in phishing and its precautionary measures, and to conduct a comprehensive study and evaluation of this research to close the gap that still exists in this area. This research focuses primarily on web - based phishing detection methods, not email - based detection methods.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AUC	Area Under the Curve
BART	Bayesian Additive Regression Trees
CART	Classification and Regression Trees
DNS	Domain Name System
FN	False Negatif
FP	False Positive
FPR	False Positive Rate
GNU	General Public License
GUI	Graphic User Interface
HTML	Hypertext Markup Language
IP	Intenet Protocol
IT	Information Technology
KNN	K-Nearest Neighbors
LR	Logistic Regression
MLBDM	Machine Learning Based Detection Method
MLP	Multi-Layer Perceptron
N	Unknown
NB	Naive Bayes
NN	Neural Networks
PD	Phishing Detection
PW	Phishing Website
PWD	Phishing Website Detection
Q1	First Quarter
Q3	Third Quarter
Q4	Fourth Quarter
RF	Random Forests
ROC	Receiver Operating Characteristics
SDLC	Software Development Life Cycle
SFH	Server Form Handler
SSL	Secure Sockets Layer
SVM	Support Vector Machines

TF–IDF	Term Frequency Inverse Document Frequency
URL	Uniform Resource Locator
WEKA	Waikato Environment for Knowledge Analysis

CHAPTER 1

INTRODUCTION

1.1 Background Overview

Phishing defined as a way of attempting to acquire information such as usernames, passwords, and credit card details by masquerading as a trustworthy entity in an electronic communication. It is a tool used by cyber criminals to steal personal information from user. The criminals will create a fake websites that look the same as the real websites.

User will get fraud by entering their confidential information such as password, bank details and account credentials into the fake websites. The fake website usually provides an embedded link to confirm the account details of the user. The criminal will then use the information provided to access the account to buy stuff, transfer money, or other damaging activities.

Phishing fraud has become the biggest threat to Internet security, according to "Chinese Network Security Report in the first half of 2011" issued by 360 SafetM, the largest security company in China. The number of phishing attacks has increased significantly in recent years, as reported by International Anti-phishing Alliance. It has become particularly urgent to find effective phishing detection methods.

1.2 Problem Statement

Internet is very useful and beneficial for everyone. The activities become online, for example, online shopping, online banking, online communication and cloud storage. However, this service is unfortunately not secure due to phishing websites.

Although there are many existing system for detecting phishing website, this systems are still unable to detect and prevent all kinds of phishing.

Moreover, existing system still have very high false alarm rated in differentiating between the phishing and normal website.

1.3 Project Objective

The main objective of this project is to detect the phishing websites. The general objectives to achieve for develop system:

i. To investigate security flaws by analyzing the state-of -the-art phishing detection system
ii. To propose a phishing detection system that analyzes website applications using machine learning
iii. To evaluate the proposed system in terms of accuracy of detection

1.4 Project Scope

For this project, it can be categorize into three scopes which are:

i. Platform
    - This application can be run on websites.
ii. Functionality
    - Computer user can be able to detect the phishing websites.

iii. User

- Every computer users (students, finance department and Government workers)

1.5 Significance

From this study, this research will find out the benefit of detecting phishing websites. There benefits that will receive are:

i. Provides organizations the safety of their websites

ii. Give banking institutions' official website more secure.

iii. Prevents internet user from get trick and have financial loss.

1.6 Thesis Content

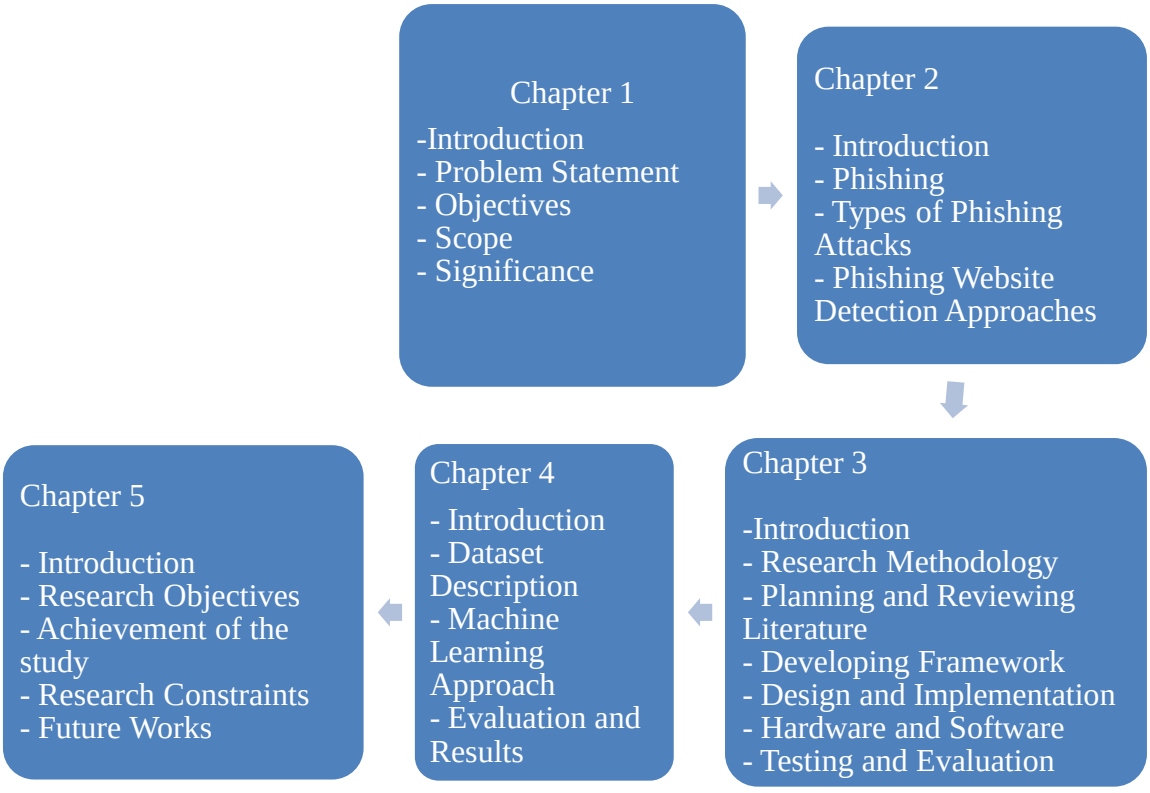

Figure 1 Summary of Each Chapter

This research will include five chapters. In Chapter one this research will discuss about introduction on this system, describe briefly information that

REFERENCES

Abu-Nimeh, S., Nappa, D., Wang, X., & Nair, S. (2007). A comparison of machine learning techniques for phishing detection. *Proceedings of the Anti-Phishing Working Groups 2nd Annual ECrime Researchers Summit on - ECrime '07*, 60–69. https://doi.org/10.1145/1299015.1299021

Amalina, F., Ali, N., Badrul, N., & Abdullah, A. (2016). Evaluation of machine learning classifiers for mobile malware detection. *Soft Computing*, 343–357. https://doi.org/10.1007/s00500-014-1511-6

APWG. (2018). *Phishing Activity Trends Report 1 Quarter*. *Most* (Vol. 1).

Chaudhry, J. A., Chaudhry, S. A., & Rittenhouse, R. G. (2016). Phishing attacks and defenses. *International Journal of Security and Its Applications*, *10*(1), 247–256. https://doi.org/10.14257/ijsia.2016.10.1.23

Chou, N., Ledesma, R., Teraguchi, Y., Mitchell, J. C., & Ca, S. (2004). Client-side defense against web-based identity theft. *Ndss*, 1–16. https://doi.org/10.1.1.65.679

Chou, T., & Pickard, J. (2018). Machine Learning based IP Network Traffic Classification using Feature Significance Analysis, *16*(3), 9–12.

Hodžić, A., & Kevrić, J. (2016). Comparison of Machine Learning Techniques. *ICESoS 2016 - Proceedings Book*, 249–256. https://doi.org/10.1109/WETICE.2011.28

Lee, J., & Kim, D. (2015). Heuristic-based Approach for Phishing Site Detection using URL Features.pdf, 131–135. Retrieved from http://eprints.ibu.edu.ba/3308/1/Adnan Hodzic Jasmin Kevric and Adem Karadag.pdf

Miyamoto, D., Hazeyama, H., & Kadobayashi, Y. (2009). An evaluation of machine learning-based methods for detection of phishing sites. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *5506 LNCS*(PART 1), 539–546. https://doi.org/10.1007/978-3-642-02490-0_66

Netcraft Ltd. (2008). Netcraft Toolbar. Retrieved from https://toolbar.netcraft.com/

Nisha, S., & Madheswari, A. N. (2016). SECURED AUTHENTICATION FOR INTERNET VOTING IN CORPORATE COMPANIES TO PREVENT PHISHING ATTACKS, *22*(1), 45–49.

Purva Sewaiwar, K. K. V. (2015). Comparative Study of Various Decision Tree Classification Algorithm Using WEKA, *9359*(10), 87–91.

Science, C. (2016). COMPARATIVE EVALUATION OF THE DIFFERENT DATA MINING TECHNIQUES, *10*(3), 233–238. https://doi.org/10.1515/ama-2016-0036

Sheng, S., Wardman, B., Warner, G., Cranor, L. F., Hong, J., & Zhang, C. (2009). An Empirical Analysis of Phishing Blacklists. *6th Conference on Email and Anti-Spam*, (March 2014). Retrieved from http://repository.cmu.edu/hcii%5Cnhttp://repository.cmu.edu/hcii/282

Thakur, R. (2015). Preprocessing and Classification of Data Analysis in Institutional System using Weka, *112*(6), 9–11.

Vanhoenshoven, F., Gonzalo, N., Falcon, R., Vanhoof, K., & Mario, K. (2016). Detecting Malicious URLs using Machine Learning Techniques. https://doi.org/10.1109/SSCI.2016.7850079

Xiang, G., Hong, J., Rose, C. P., & Cranor, L. (2011). Cantina+. *ACM Transactions on Information and System Security*, *14*(2), 1–28. https://doi.org/10.1145/2019599.2019606